Review of Quantitative Finance and Accounting
https://doi.org/10.1007/s11156-026-01500-1

ORIGINAL RESEARCH

A graphical procedure for equity premium and stock return prediction: Monte Carlo evidence

Neil M. Kellard[1] · **Fotios I. Papadimitriou**[2]

Received: 2 October 2024 / Accepted: 13 January 2026

Abstract
This paper assesses the robustness of the popular Goyal and Welch graphical procedure which has been extensively used in the literature to evaluate the performance of predictive models for stock returns among other contexts. To this end, we simulate the graphical diagnostic and construct a sign-based test allowing us to examine its behaviour under various sample sizes, data generating processes and levels of correlation. Our simulations reveal that correlation does have an effect on the graph in smaller samples but the technique is quite robust when sufficiently large data sets are employed. Moreover, we demonstrate that the graphical diagnostic is generally well-sized and yields a satisfactory power performance in most cases. This result holds also under the assumption of heteroskedasticity. Overall, our analysis suggests that the graphical diagnostic can be an important complement to the more conventional methods seeking to assess out-of-sample predictive ability.

Keywords Stock return predictability · Monte carlo simulation · Out-of-sample prediction · Recursive forecasts · Dividend ratios

JEL classification C15 · C22 · C53 · G17

1 Introduction

During the last four decades, a vast literature has well documented the ongoing debate regarding whether stock market returns are predictable. The sheer variety of testing procedures and complicated methodologies that have been proposed as well as the numerous markets and sample periods that have been examined, may give a clear explanation of why

✉ Fotios I. Papadimitriou
fotios.papadimitriou@abdn.ac.uk

Neil M. Kellard
nkellard@essex.ac.uk

[1] Essex Business School, University of Essex, Colchester CO4 3SQ, UK

[2] Business School, University of Aberdeen, Aberdeen AB24 3QY, Scotland, UK

Published online: 08 April 2026

it is difficult to achieve a general consensus.[1] On the one hand, the extant literature provides ample evidence of both in-sample and out-of-sample stock return predictability by means of various financial and macroeconomic variables or technical indicators (see, *inter alia*, Rozeff 1984; Campbell and Shiller 1988a, b; Fama and French 1988; Lamont 1998; Pontiff and Schall 1998; Lettau and Ludvigson 2001; McMillan 2003; Pesaran and Timmermann 2000; Rapach and Wohar 2006; Campbell and Thompson 2008; Kellard et al. 2010; Rapach et al. 2013; Jordan et al. 2014; Neely et al. 2014; Charles et al. 2017; Kuntz 2020; Tsiakas et al. 2020).

On the other hand, there are several studies that find no predictable components in stock returns. For example, Bossaerts and Hillion (1999) suggest that even the best prediction models have weak out-of-sample predictive power while Goyal and Welch (2003) show that the popular dividend ratios are not useful predictors of the US equity premium. In a comprehensive study, Welch and Goyal (2008) demonstrate that virtually all possible predictive variables perform poorly against the historical moving average benchmark model and stress that this result is a systemic problem not restricted to any decade.[2] The case of weak evidence for return predictability is also supported by Ang and Bekaert (2007), Choi et al. (2016) and Goyal et al. (2024), among others.

Furthermore, a strand of literature is related to the potential econometric problems of stock return predictability. Small sample biases in predictive regressions, overlapping observations and highly persistent predictor variables are the focal point of various studies which have re-examined and criticized the evidence of predictability (see, Goetzmann and Jorion 1993; Nelson and Kim 1993; Ferson et al. 2003; Goyal et al. 2024). As a result, a growing body of research work has proposed some new testing procedures which involve bias-corrected estimators in predictive regressions for valid inference (e.g., Lewellen 2004; Amihud and Hurvich 2004; Amihud et al. 2004; Campbell and Yogo 2006; Hjalmarsson 2011; Kostakis et al. 2015; Harvey et al. 2023).

A very important issue within the return predictability literature is the fact that finding in-sample statistical significance does not necessarily mean that the employed variables will also exhibit a successful predictive performance out-of-sample. Out-of-sample tests are typically believed to provide an approach to mitigate the effects of data mining and as such, they are often used for a more robust evaluation of competing forecasting models.[3] More importantly, they are of particular interest to investors who aim to improve their investment strategies by relying on real-time predictions of stock returns. Within this context, Goyal and Welch (2003) (henceforth GW) introduce a recursive residuals (out-of-sample) graphical procedure for equity premium and stock return prediction and firmly suggest that future papers adopt their approach when investigating the market timing ability of different predictive variables. An interesting aspect of this methodology stems from the more dynamic identification of predictability. One can clearly observe the time periods where the predictive variable succeeds or fails in predicting the equity premium out-of-sample. Hence,

[1] For a comprehensive overview of the return predictability literature, see Rapach and Zhou (2022).

[2] Campbell and Thompson (2008) show that once sensible restrictions are imposed on the signs of the coefficients and stock returns when constructing out-of-sample forecasts, the out-of-sample explanatory power can be economically meaningful for investors.

[3] Interestingly, a few papers have challenged this "conventional wisdom". For example, Inoue and Killian (2004) and Rapach and Wohar (2006) support the view that if appropriate tests are employed, in-sample and out-of-sample tests are equally reliable.

it makes it easy to understand the relative performance of the competing forecasting models. Although graphing recursive residuals is a relatively simple approach, ignoring its use or findings may leave useful information regarding predictability uncovered. Therefore, it is possible that when conventional summary measures indicate no predictability, the graphical procedure may suggest otherwise and reveal concealed aspects of forecastability.

The usefulness and intuitive appeal of the graphical diagnostic suggested by GW has led to its extensive use in the literature by numerous studies within various settings of predictability. For example, it has been employed to explore the predictive ability of different models for stock returns (e.g., Robertson and Wright 2006; Welch and Goyal 2008; Schrimpf 2010; Kellard et al. 2010; Rapach et al. 2010; Andriosopoulos et al. 2014; Pettenuzzo and Ravazzolo 2016; Algaba and Boudt 2017; Lawrenz and Zorn 2017; Lima and Meng 2017; Zakamulin 2017; Yin 2019; Yin et al. 2019; Félix et al. 2020; Kuntz 2020; Stöckl and Kaiser 2021; Yin 2021; Li et al. 2022; Sakkas and Tessaromatis 2022; Yin 2022; Alexandridis et al. 2023; Deng et al. 2024; Liu et al. 2024; Nygaard and Sørensen 2024; Park et al. 2024; Wang et al. 2024; Chen et al. 2025), stock market volatility (e.g., Chen et al. 2016; Xie 2019; Ciner 2025), oil price returns (e.g., Zhang et al. 2019), oil and gas volatility (e.g., Xu and Lien 2022; Luo et al. 2024), bond excess returns (e.g., Yin and Yang 2024), bond yields (e.g., Caldeira et al. 2016; Caldeira and Torrent 2017), exchange rates (e.g., Foroni et al. 2018), gross domestic product (e.g., Bjørnland et al. 2017), inflation (e.g., Hong et al. 2025), interbank rates (e.g., Monticini and Ravazzolo 2014), leverage (e.g., Amini et al. 2021), credit default swaps (e.g., Procasky and Yin 2022; Procasky and Yin 2023), commodity returns (Angelidis et al. 2025; Gao et al. 2025), and cryptocurrency returns (e.g., Bennett et al. 2024). Hence, it is clearly established as a very popular technique among studies that seek to assess predictive performance within diverse frameworks.

This paper extends the work of GW and contributes to the return predictability literature by employing Monte Carlo simulations to assess the robustness of their proposed graphical procedure. This is an important issue given its appealing nature and widespread use in various contexts which involve the comparison of competing forecasting models. However, little is known about its finite-sample properties and behaviour. Our aim is to fill this gap and provide new evidence that will allow us to evaluate the usefulness of this diagnostic test as a complementary measure for out-of-sample predictive ability. To that end, we construct a sign-based test which is derived from the algebraic representation of the procedure.[4] Therefore, our analysis facilitates a "formal" comparison between the graphical diagnostic and the more conventional tests of equal predictive accuracy. To our knowledge, this is a unique concept that has not been applied in the literature. We offer evidence with respect to various sample sizes, empirically relevant proportions of out-of-sample to in-sample observations for each case and we also account for different levels of correlation between the innovations in the returns and the predictive variable processes.[5] The returns series are generated based on normally distributed errors but as a robustness check we extend the analysis to consider the case of GARCH(1,1) innovations.

The first step in our methodology is to construct simulated out-of-sample forecasts from two models: the unconditional historical moving average model and the conditional divi-

[4] In a different context, theoretical work by Christoffersen and Diebold (2006) suggests that volatility dependence produces sign dependence, and hence forecastability, as long as expected returns are non-zero.

[5] The latter is represented by the dividend-price ratio which is one of the most prominent candidates to predict stock returns.

dend-price ratio model. Subsequently, we are able to simulate the graphical diagnostic by calculating its algebraic representation at each replication. We then employ a sign-based test and focus on the percentage of positive points of the graphical procedure which is an indication of predictive ability. Empirical critical values for this expression are derived, allowing us to explore the finite-sample size and power properties of the graphical method.

Additionally, in each Monte Carlo draw we compute the classic Diebold and Mariano (1995) test statistic of equal predictive accuracy, one of its modifications proposed by Harvey et al. (1997) and a subsequent influential test developed by McCracken (2007). Our paper provides empirical evidence and discusses the size and power properties of these statistics.[6] The Diebold and Mariano (1995) test and its modified version by Harvey et al. (1997) are widely used in the literature until today (for recent applications in various contexts, see, Gkillas et al. 2021; Adediran and Swaray 2023; Asgharian et al. 2023; Basistha 2023; Ellwanger and Snudden 2023; Gao et al. 2023; Campisi et al. 2024; Lan et al. 2024; Salisu et al. 2024; Yang et al. 2024; Hong et al. 2025; Labonne 2025). On the other hand, the statistic of McCracken (2007) is generally found to be more powerful and to yield better size properties in finite samples (see also, Clark and McCracken 2009).[7]

Turning to our preliminary results, we find that a high correlation between the errors in the excess returns and the dividend-price ratio processes has some effect on the graphical procedure in small samples. However, this effect diminishes as the sample size increases and the graphical diagnostic appears to be quite robust. This result holds under different scenarios (e.g., sample sizes, proportions of out-of-sample to in-sample observations or under the assumption of heteroskedasticity in the errors).

Our Monte Carlo analysis then focuses on the finite-sample size properties of the graphical procedure. Specifically, we report estimates of the probability of making a type I error under the null hypothesis of equal predictive accuracy between the two competing models. As expected, we find some evidence of over-rejecting the null in small samples (especially under the presence of high correlation). This is also true for the more conventional statistics under consideration. However, as the sample size increases, all statistics appear to be well-sized and the effect of correlation disappears in most cases. These results remain consistent when we consider two different specifications of GARCH(1,1) innovations to generate the returns series.

Finally, we explore the empirical power properties of the graphical procedure along with the more conventional tests of equal predictive accuracy. Under different scenarios, we demonstrate that for all statistics there is a significant gain in power as the sample size increases and they exhibit a satisfactory performance. In a strictly statistical sense and in line with earlier work, we find that the formal test suggested by McCracken (2007) generally yields the highest power.

Overall, our study reveals that the graphical procedure of GW exhibits satisfactory and stable finite-sample properties and it is well-sized under most scenarios within our Monte Carlo experiment. Hence, in conjunction with its dynamic and insightful nature, we suggest that it can act as a useful complementary tool for assessing the relative predictive performance

[6] Note that our aim is not to provide an exhaustive comparison between various well-known statistics of equal predictive accuracy or their extensions. Our main focus is on the finite-sample behaviour of the graphical procedure of GW as a complementary diagnostic test.

[7] For recent evidence employing this statistic, see Berisha et al. (2021), Bouri et al. (2021), Costantini and Kunst (2021), Stauskas and Westerlund (2022), Su et al. (2022) and Gupta et al. (2023).

of different models. This is of particular interest to academics and investors who rely on real-time data to construct forecasts and to form trading strategies. Our findings also indicate that caution is warranted when very small samples are considered, given that the corresponding results may be susceptible to some biases in such cases (especially under the presence of high correlation). However, this is something that may affect all other well-known tests of predictive accuracy and it is not limited only to the use of the graphical diagnostic.

The remainder of the paper is organized as follows. Section 2 describes the methodology and carries out the Monte Carlo experiments. Section 3 discusses the empirical results and a final section concludes.

2 Methodology

2.1 Simulation design

2.1.1 Gaussian errors

Let r_t denote the excess stock return in period $t = 1, ..., T$ and let x_t denote the corresponding value of the predictive variable of interest such as the dividend-price ratio. We rely on the standard data generating process (DGP) most often found in the literature (see, *inter alia*, Campbell and Yogo 2006; Clark and West 2006; Rapach and Wohar 2006) and we postulate that the data are generated according to:[8]

$$r_t = a + bx_{t-1} + u_{1t}, \tag{1}$$

$$x_t = \mu + \rho\, x_{t-1} + u_{2t}, \tag{2}$$

where the joint innovations $u_t = (u_{1t}, u_{2t})'$ are independently and identically distributed (i.i.d.), normal with unit variance and correlation δ. Consequently, we set $u_{1t} = \delta\, u_{2t} + \sqrt{1-\delta^2}v_t, v_t \sim N(0,1)$ and we let the correlation vary between 0 and -0.9. The negative correlation between the innovations to excess returns and the dividend-price ratio is a common assumption in the return predictability literature which is empirically supported by previous studies (e.g., Engstrom 2003; Campbell and Yogo 2006; Hjalmarsson 2011). The intercepts a and μ are set to 0.05 and -3.31 respectively,[9] while the autoregressive root ρ is set equal to 0.93.[10] In experiments evaluating size (i.e. under the null of no predictability), $b = 0$ in Eq. (1). On the other hand, $b \neq 0$ in experiments evaluating power (i.e. under the alternative).

2.1.2 GARCH(1,1) errors

In this section, we extend the DGP described above to consider data with conditional heteroskedasticity and the fat-tails features that are often thought to characterize financial data.

[8] All simulations are carried out in Ox (see, https://www.doornik.com/).

[9] Our results are invariant to these coefficients (both under the null and the alternative hypotheses).

[10] The model used for the dividend-price ratio was derived by fitting an AR(1) process to UK monthly data taken from the UK FTSE All-Share index (see, Kellard et al. 2010).

For that purpose, we generate GARCH(1,1) innovations for the returns series which are driven either by the standard normal distribution or by the Student's *t*-distribution with five degrees of freedom ($t(5)$).[11] In particular, the GARCH(1,1) model is $u_{1t} = z_t \sigma_t$ where $z_t \sim i.i.d.N(0,1)$ or $z_t \sim t(5)$ and $\sigma_t^2 = \lambda_0 + \lambda_1 u_{1t-1}^2 + \lambda_2 \sigma_{t-1}^2$, where λ_0, λ_1 and λ_2 are constants.[12] Provided that the condition $\lambda_1 + \lambda_2 < 1$ is satisfied, the u_{1t} series is covariance stationary. Evidence from stock market data suggests that $\lambda_1 + \lambda_2$ as well as λ_2 are close to one. He and Teräsvirta (1999) show that the unconditional fourth moment of u_{1t} exists for GARCH(1,1) models if and only if $\lambda_2^2 + 2\lambda_1\lambda_2 E|z_t|^2 + \lambda_1^2 E|z_t|^4 < 1$. We set $\lambda_0 = 0.06$, $\lambda_1 = 0.05$ and $\lambda_2 = 0.89$. With this parameterization, the He and Teräsvirta (1999) condition for the existence of the fourth moment is satisfied and also, it is ensured that the unconditional variance of u_{1t} is the same as in the homoskedastic case (i.e. equal to one).

2.2 Predictive regressions and out-of-sample simulated forecasts

Typically, the evaluation of in-sample predictability involves the estimation of the following regression model:

$$r_t = \alpha + \beta x_{t-1} + \epsilon_t, \tag{3}$$

where r_t is the equity premium and x_{t-1} is the lagged predictive variable of interest such as the dividend-price ratio. The predictive ability of x_{t-1} is assessed by examining the *t*-statistic corresponding to $\hat{\beta}$, the OLS estimator of β in Eq. (3), as well as the goodness of fit measure, R^2. The null hypothesis of no predictability implies that $\beta = 0$.

As in GW, we focus on a real-time market investor and use only then-available data in our simulations. Consequently, the question of interest is how the conditional dividend ratio model would perform out-of-sample, when compared against the unconditional historical equity premium model (the prevailing historical moving average). For both models, we use the recursive scheme to predict one-step-ahead equity premia. First, the total sample of T observations is divided into in-sample and out-of-sample proportions. The in-sample observations span from 1 to R. Letting P denote the number of one-step-ahead predictions, the out-of-sample observations span $R+1$ through $R+P$. Consequently, regarding the dividend model, Eq. (3) is estimated repeatedly via OLS and forecasts are updated as additional data become available in each period. On the other hand, the prevailing up-to-date equity premium average is used to predict the next period's equity premium.

Results are reported for sample sizes varying between 80 and 1600 observations and we also consider empirically relevant combinations of in-sample and out-of-sample proportions for each sample. In particular, we let the ratio $\pi = \frac{P}{R}$ take values between 0.4 and 2.0.

The next sub-section discusses the conventional tests of equal predictive accuracy which are employed in this study.

[11] Respectively, the errors in the x_t process will be standard normal or a linear combination of underlying innovations drawn from the $t(5)$ distribution.

[12] Now we have: $corr(u_{2t}, z_t) = \delta$ (see, for example, Clark and West 2006).

2.3 Conventional tests for comparing predictive accuracy

The first test we consider is the classic Diebold and Mariano (1995) statistic (henceforth DM) which assumes equal predictive ability between two competing models. The approach of DM can be summarized as follows.

Consider *h*-step-ahead forecasts in a P row vector and let $\hat{y}_{1,t}$ and $\hat{y}_{2,t}$ denote the forecasts of y_t obtained from two different models; now let $g(\epsilon_{it}), i = 1,2$ be some arbitrary loss function where ϵ_{it} is the corresponding forecast error (that is, $\epsilon_{it} = y_t - \hat{y}_{i,t}, i = 1,2$). The null hypothesis of equal predictive accuracy can be expressed as: $H_0 : E[g(\epsilon_{1t})] = E[g(\epsilon_{2t})]$ or equivalently $H_0 : E[d_t] = 0$ where $d_t = g(\epsilon_{1t}) - g(\epsilon_{2t})$.[13] DM show that the asymptotic distribution of the sample mean loss differential $\bar{d} = P^{-1}\sum_{t=1}^{P}[g(\epsilon_{1t}) - g(\epsilon_{2t})] = P^{-1}\sum_{t=1}^{P} d_t$ is given by $\sqrt{P}(\bar{d} - \mu) \xrightarrow{d} N(0, V(\bar{d}))$ where $V(\bar{d})$ is the variance of $\bar{d}$. It can be shown that a consistent estimator of the asymptotic variance $V(\bar{d})$ is given by:

$$\hat{V}(\bar{d}) = \frac{1}{P}[\hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k], \tag{4}$$

where $\hat{\gamma}_k$ is an estimate of the k^{th} autocovariance of d_t, calculated as:

$$\hat{\gamma}_k = \frac{1}{P}\sum_{t=k+1}^{P}(d_t - \bar{d})(d_{t-k} - \bar{d}), \tag{5}$$

The DM test statistic is then expressed as follows:

$$DM = \frac{P^{-1}\sum_{t=1}^{P}[g(\epsilon_{1t}) - g(\epsilon_{2t})]}{\sqrt{\hat{V}(\bar{d})}}, \tag{6}$$

Under the null, the DM statistic follows asymptotically the standard normal distribution. For this to be true however, some conditions must hold.

For example, when applied to non-nested models, West (1996) shows that the DM statistic can be asymptotically standard normal. However, McCracken (2007) shows that for forecasts from nested models it has a non-standard limiting distribution and he provides tables with asymptotically valid critical values. Thus, in our Monte Carlo analysis we use the asymptotic critical values tabulated by McCracken (2007).

Furthermore, in each draw we employ a modified version of the original DM statistic (henceforth MDM), proposed by Harvey et al. (1997), that corrects for size distortions in small samples. Harvey et al. (1997) suggest that critical values from the Student's *t*-distribution (with $P - 1$ degrees of freedom) are more appropriate when the MDM statistic is considered. For an *h*-step-ahead forecast in a P row forecast vector, the MDM is given by:

[13] When comparing mean squared errors (MSE) between two nested models as we do, the appropriate loss function is of the form: $g(\epsilon_{it}) = \epsilon_{it}^2, i = 1,2$ where $i = 1$ stands for the unconditional model and $i = 2$ corresponds to the conditional dividend model.

$$MDM = \sqrt{\frac{P+1-2h+P^{-1}h(h-1)}{P}}\, DM, \tag{7}$$

For $h=1$, we have that $MDM = \sqrt{(P-1)/P}\, DM$. Again, while theory suggests that the MDM follows a *t*-distribution for non-nested models, it has a non-standard limiting distribution when comparing forecasts between nested models. Hence, in our Monte Carlo analysis we use the numerical estimates of the asymptotic critical values provided by McCracken (2007).

Extending the earlier empirical work, McCracken (2007) develops an out-of-sample *F*-type test which is designed to test for equal predictive accuracy based on the mean squared error criterion (MSE). He shows that the new statistic converges in distribution to a function of stochastic integrals of quadratics of Brownian motion and derives asymptotic critical values for valid inference. Compared to the other alternatives, the proposed statistic is found to be more powerful and to have better size properties in extensive Monte Carlo experiments. Denoting it by MCCR, its algebraic representation can be expressed as follows:

$$MCCR = P\frac{P^{-1}\sum_{t=1}^{P}\left[g\left(\epsilon_{1t}\right) - g\left(\epsilon_{2t}\right)\right]}{\widehat{c}}, \tag{8}$$

where $\epsilon_{it}, i=1,2$ denote the forecast errors from the two competing models and $g\left(\epsilon_{it}\right), i=1,2$ is the relevant quadratic loss function as defined earlier. Finally, $\widehat{c}$ converges in probability to a certain normalizing constant. In our case, $\widehat{c}$ is defined as $P^{-1}\sum_{t=1}^{P} g\left(\epsilon_{2t}\right) = P^{-1}\sum_{t=1}^{P}\epsilon_{2t}^{2}$. Under this setup, this statistic takes the form of the standard *F*-test but adapted to an out-of-sample context (see, McCracken 2007). It is important to note here that in the case of nested models the tests are one-sided. The null hypothesis of equal predictive accuracy is tested against the alternative that the unconditional model produces less accurate forecasts compared to the conditional dividend ratio model (i.e. it has greater mean squared forecast error).

Next, we present the graphical procedure and describe the simulation approach we follow with the purpose of exploring its properties.

2.4 A graphical diagnostic test for out-of-sample predictive ability

2.4.1 Background

As mentioned earlier, GW propose a graphical method which they claim can act as a powerful diagnostic test for equity premium and stock return prediction. The procedure consists of plotting (against time) the cumulative sum-squared error from the unconditional model minus the cumulative sum-squared error from the dividend ratio model (denoted by $Net-SSE_T$) and is given by:

$$Net-SSE_T = \sum_{t=T-R+1}^{T}[SE_t^{Prevailing\ Mean} - SE_t^{Dividend\ Model}], \tag{9}$$

where T is the total sample size, R is the in-sample size and SE_t is the squared out-of-sample prediction error in observation t. The prevailing up-to-date equity premium aver-

age gives the estimate of the unconditional model. The conditional prediction errors of the dividend model are derived from recursive regressions with the dividend-price ratio being the sole predictor of the next period's equity premium (see Sect. 2.2). It is evident from the above expression that a positive value implies the superior performance of the dividend ratio model against the historical moving average model so far. Additionally, a positive slope suggests that the dividend model had lower forecasting error in a given period.

2.4.2 Simulating the graphical procedure

As noted in the introduction, the aim of this paper is to provide new evidence and assess, via Monte Carlo simulations, the robustness and usefulness of GW's methodology in terms of equity premium and stock return prediction. For that purpose, we construct a sign test which is based on the graphical procedure. At each replication in our simulations, the algebraic expression shown in Eq. (9) is estimated. In order to evaluate the performance of the conditional model we focus on the percentage of the values of $Net - SSE_T$ which are positive in the out-of-sample period. Therefore, at each replication i, considering n_1 sample sizes T_j, and n_2 different proportions of out-of-sample to in-sample observations (π_k) for each sample, we compute:

$$q_i^{jk} = \frac{Q_i^{jk}}{P^{jk}} \times 100, Q_i^{jk} = \sum_{T_j - R^{jk}+1}^{T_j} I\left[(Net - SSE_{T_j})_i > 0\right], i = 1, ..., N, j = 1, ..., n_1, k = 1, ..., n_2, \quad (10)$$

where at replication i, Q_i^{jk} is the number of positive out-of-sample observations of the graph with respect to sample size T_j and ratio π_k, out of a total of P^{jk} forecasts for each case. Furthermore, R^{jk} denotes the number of in-sample observations and I is the indicator function. For the above cases, our first results report the average of all percentages (under the null) considering the number of repetitions of the experiment (we call it $\bar{t}_{jk}$):

$$\bar{t}_{jk} = \frac{\sum_{i=1}^{N} q_i^{jk}}{N}, j = 1, ..., n_1, k = 1, ..., n_2, \quad (11)$$

where N is the number of Monte Carlo replications. The above experiment is also carried out under different levels of correlation δ between the returns and the dividend-price ratio processes.

2.4.3 Exploring the size and power properties of the graphical procedure

This section delves deeper into the properties of q_i^{jk} which is described in the previous section. As mentioned above, in each draw i, q_i^{jk} represents the percentage of positive points (out-of-sample) of the diagnostic test, with j and k as defined in Eq. (10). The graphical procedure is not a standard measure of predictive ability with any kind of known behaviour. Therefore, we derive empirical critical values for q_i^{jk} and we perform one-sided tests at the 5% level of significance. This approach enables us to conduct size and power simulations. Additionally, it facilitates a more "formal" comparison between the graphical diagnostic and the conventional tests of equal predictive accuracy.

We generate a large sample of 6,000 observations (i.e. $j = 1$) and, considering different out-of-sample to in-sample proportions (π_k), we repeat the experiment 50,000 times,

Table 1 Empirical critical values for the graphical procedure

$\pi = \frac{P}{R}$	99th Percentile	95th Percentile	90th Percentile
0.1	1.000	0.982	0.932
0.2	0.999	0.973	0.908
0.4	0.998	0.956	0.864
0.6	0.997	0.944	0.835
0.8	0.996	0.926	0.795
1.0	0.994	0.907	0.772
1.2	0.992	0.899	0.743
1.4	0.991	0.877	0.721
1.6	0.989	0.869	0.699
1.8	0.988	0.859	0.688
2.0	0.985	0.840	0.667
3.0	0.977	0.782	0.601

Explanation: This table tabulates the calculated empirical critical values at the usual levels of significance, for the sign-based test inspired by the graphical procedure of Goyal and Welch (2003). The 99th, 95th and 90th percentiles were derived from an experiment which uses a sample of 6,000 observations. The number of repetitions is 50,000. Section 2.4.2 describes the simulation approach and Section 2.4.3 explains how we obtain the critical values. The ratio of out-of-sample (*P*) to in-sample (*R*) observations is denoted by π. Table 1 shows the results with respect to various values of π

calculating q_i^{jk} at each replication.[14] In this way, we obtain an empirical distribution for q_i^{jk}. For each case, the 95th percentile of the empirical distribution is chosen as the critical value at the 5% level.[15] Table 1 tabulates the computed critical values for various values of π_k ($k = 1, ..., 12$) and with respect to the levels of significance 1%, 5% and 10%.[16]

Next, we set up new experiments for all sample sizes under consideration and, in each draw, we compare the computed q_i^{jk} against the simulated critical values, obtaining the empirical size and power of the graphical diagnostic. For a more direct comparison, size and power results are also reported for the well-known statistics described in Sect. 2.3.

3 Results

3.1 Simulation results for the graphical method

Table 2 presents preliminary Monte Carlo evidence (under the null of no predictability) for the recursive residuals diagnostic test that is described in Sect. 2.4. Considering 25,000 simulated graphs (i.e. the number of repetitions of the experiment is 25,000) we report, in terms of average percentages, the number of point estimates of $Net - SSE_T$ which are

[14] This experiment is carried out under the null of no predictability. Also, correlation δ is set to zero since it does not seem to have an effect on q_i^{jk} when larger samples are employed.

[15] Analogously, we obtain the corresponding critical values with respect to the 1% and 10% significance levels.

[16] Note that, henceforth, the subscripts or superscripts j and k are omitted in the tables for the sake of simplicity.

Table 2 Simulation results for $\bar{t}$ under the null of no predictability

	DGP	Panel A Normal errors			Panel B GARCH(1,1)-normal errors			Panel C GARCH(1,1)- t (5) errors		
		δ			δ			δ		
Sample size	$\pi = \frac{P}{R}$	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	−0.9
T = 80	0.4	34.0	35.6	40.1	34.0	35.6	40.4	33.3	36.0	40.3
	0.6	30.9	32.2	37.2	31.4	32.1	36.8	31.1	33.6	37.6
	1.0	27.1	28.4	33.2	26.6	29.2	32.0	26.7	29.3	33.9
	2.0	20.9	22.5	27.1	20.7	22.6	27.4	21.7	23.6	26.9
T = 400	0.4	33.4	34.4	35.8	33.5	33.9	35.9	33.3	34.6	36.0
	0.6	30.9	31.5	32.9	30.3	31.6	33.9	30.1	31.6	33.7
	1.0	27.5	27.5	29.9	27.1	27.8	28.6	27.3	28.1	29.8
	2.0	22.2	23.1	24.4	22.3	23.0	24.7	21.6	23.6	24.9
T = 1000	0.4	33.3	33.4	34.4	32.7	34.0	34.0	33.4	33.6	33.5
	0.6	30.5	31.3	31.8	30.3	30.5	31.9	30.0	30.4	31.4
	1.0	27.0	27.2	28.1	27.5	28.2	27.9	27.5	27.8	28.4
	2.0	22.7	22.8	23.7	22.7	22.8	22.8	22.2	23.3	23.7
T = 1600	0.4	33.2	33.0	33.4	32.9	33.6	33.7	33.1	33.8	34.6
	0.6	30.3	30.9	30.8	31.2	31.0	30.7	30.0	30.6	31.9
	1.0	27.2	27.4	28.2	26.8	27.5	28.1	27.6	27.2	28.0
	2.0	22.3	22.8	23.2	22.5	22.3	22.9	22.8	22.8	23.4

Explanation: This table presents preliminary simulation results for the graphical procedure of Goyal and Welch (2003) that was described in Section 2.4.1. The graphical diagnostic is simulated over 25,000 replications. At each replication i, we calculate the fraction of the point estimates of the graph which are positive in the out-of-sample period. This table tabulates the average percentages of positive points considering the total number of replications, denoted by t (see, section 2.4.2 for details). We consider different proportions of out-of-sample (P) to in-sample (R) observations, denoted by π, as well as different levels of correlation (δ) between the errors of the returns and dividend-price ratio processes. Panel A reports results with respect to the returns data generating process (DGP) that is based on Gaussian errors while Panels B and C show the results with respect to the DGP that uses GARCH(1,1)-normal errors and GARCH(1,1)-t errors with five degrees of freedom, respectively (see Section 2.1 for a detailed description)

positive in the out-of-sample period. This is expressed in Eq. (11) (see Sect. 2.4.2) and it is denoted by $\bar{t}_{jk}$ (or $\bar{t}$ for simplicity). Table 2 shows the results for sample sizes of 80, 400, 1000 and 1600 observations and for proportions of out-of-sample to in-sample observations (π) which take the values 0.4, 0.6, 1.0 and 2.0. Finally, we allow for the negative correlation δ that is often found between the errors in the excess returns and dividend-price ratio processes, a quite common assumption in empirical work (e.g., Campbell and Yogo 2006; Hjalmarsson 2011).

With respect to the case of Gaussian errors (see Sect. 2.1.1), the results in Panel A show that the presence of correlation affects the graphical procedure in small samples but the diagnostic appears to be quite robust as the sample size increases. For example, when the sample size is 80, π is 0.6 and there is no correlation (i.e. $\delta = 0$), our simulations suggest that we would expect the conditional model to outperform about 30.9% of the time. However, the corresponding percentage rises to 37.2% when correlation is set to -0.9. On the other hand, the larger data sets we employ, the less correlation affects the results. Take for example the case of 1600 observations, with π being 0.6: $\bar{t}$ would rise from 30.3% with no correlation, to 30.8% when correlation is set to -0.9.

Another factor that seems to play an important role is the way we split the sample size so as to construct out-of-sample forecasts. For all cases, as π increases (i.e. fewer observations are held to produce forecasts) the average percentages drop significantly. For instance, if we employ a sample of 400 observations with $\pi = 0.4$ and $\delta = -0.5$, we would expect the dividend model to outperform the historical moving average model 34.4% of the time. On the contrary, if we decrease the in-sample size (R) so as π increases to 2.0, the corresponding percentage drops to 23.1%. This result implies that it is more difficult to detect any predictive ability when longer out-of-sample periods are used.

Results for $\bar{t}$ are almost identical when we employ GARCH(1,1) errors to construct the returns series and thus, there is no need to further comment on those (see Table 2, Panels B and C). This is clearly a positive sign for the robustness of the graphical diagnostic since it exhibits similar behaviour under the assumption of heteroskedasticity in the errors, to the behaviour under normality.

3.2 Simulation results: Size

In order to assess the finite-sample performance of the graphical diagnostic (denoted by GW in the tables) together with the performance of the more conventional statistics (see Sect. 2.3), we run simulations and report estimates of the probability of making a type I error.[17] The null hypothesis assumes equal predictive accuracy between the two competing forecasting models. As mentioned in the methodology, all tests are one-sided. The one-sided alternative hypothesis implies that the unconditional model has a higher forecasting error than the dividend model. Table 3 tabulates the empirical rejection rates at the 5% nominal size.[18] The number of simulations is 25,000. Critical values for the conventional tests are taken from McCracken (2007) while for the graphical diagnostic we use our own empirical critical values, as explained in Sect. 2.4.3 and presented in Table 1.

We start the discussion from Panel A which is related to the data generating process that uses Gaussian errors to construct the returns series. Looking at the small sample of 80 observations, we find that, in the absence of correlation, all statistics are relatively well-sized although they are oversized when π is 0.4 and slightly undersized when π is 2.0. For all levels of correlation, the statistics exhibit their worst performance when π is smallest and equal to 0.4. The graph-based sign test rejects the null between 8% and 10.9%, the MDM statistic between 6.7% and 10.5% while the MCCR statistic performs better with rejection rates between 5.9% and 8.2%. Regarding the other cases of π's (between 0.6 and 2.0), when correlation is modest (i.e. $\delta = -0.5$), the graphical diagnostic retains good size properties with rejection probabilities ranging from 5.6% (Panel A, $\pi = 2.0$) to 5.9% (Panel A, $\pi = 0.6,\ 1.0$), outperforming the MDM statistic which has actual sizes between 4.7% (Panel A, $\pi = 2.0$) and 6.4% (Panel A, $\pi = 1.0$). The MCCR statistic exhibits the best size properties with empirical sizes ranging from 4.9% (Panel A, $\pi = 2.0$) to 5.6% (Panel A, $\pi = 1.0$). However, when correlation is strongly negative (i.e. $\delta = -0.9$), both the graphical procedure and the MDM statistic appear more oversized (reaching actual sizes of 8.9% and 8.6%, respectively) while the MCCR statistic

[17] By definition, the probability of committing a type I error is the probability of rejecting the null when it is true.

[18] Results with respect to the DM statistic are not reported since in larger samples they are almost identical to the MDM statistic. The latter however, performs slightly better when the sample size is 80.

is found relatively more reliable (with rejection probabilities between 6.5% and 7.3%). Furthermore, the size distortions of the tests fall as the number of post-sample predictions P rises and π becomes larger.

As the sample size increases, all statistics appear well-sized and the effect of correlation seems to diminish in most cases. For example, when the sample size is 1000 and $\pi = 2.0$, GW rejects the null between 5.1% and 5.7%. In general, the finite-sample performance of the graphical diagnostic is quite satisfactory as it exhibits similar size properties to the more conventional tests.

The results follow a similar pattern if we extend the DGP to consider the case of GARCH(1,1)-normal errors and the MCCR statistic yields better size properties in small data sets. Still, all tests are well-sized when sufficiently large data sets are employed (Table 3, Panel B).

On the other hand, when we consider data generated from a fat tailed i.i.d. distribution (Table 3, Panel C), we observe some differences compared to the normal or to the GARCH(1,1)-normal case. Although the MCCR statistic is still better for the small sample of 80 observations, its empirical size has now increased compared to the homoskedastic case. Additionally, its performance has relatively worsened in some cases compared to the other two statistics. For example, when we employ a sample of 1600 observations and correlation is set to -0.9, its rejection probabilities are 7.0%, 6.2%, 6.6% and 6.1% for π being equal to 0.4, 0.6, 1.0 and 2.0, respectively. The corresponding rejection rates in the case of Gaussian errors were 5.5%, 4.7%, 5.3% and 5.0% (Table 3, Panel A). On the contrary, the graphical diagnostic has improved its size properties in three out of four cases and yields empirical sizes between 5.3% and 6.0%. The MDM statistic in this case rejects the null between 4.9% and 6.1%. It should be noted however that, again, all statistics are well-sized in most cases and the results do not change significantly when we employ GARCH(1,1) errors to generate the returns series.

3.3 Simulation results: Power

In this section, we report the empirical power of the graphical procedure and the conventional tests of equal predictive accuracy at the 5% nominal level.[19] Panels A, B and C of Table 4 show the results for the cases of normal, GARCH(1,1)-normal and GARCH(1,1)-$t(5)$ errors that were presented in Sect. 2.1. The parameter values of b in Eq. (1) are chosen to be 0.01, 0.02 and 0.05.[20] For brevity, Table 4 tabulates power results only for sample sizes of 80 and 1600 observations (i.e. the two extreme considered cases in this study). The number of replications is 25,000.

With respect to the DGP with normal errors (Table 4, Panel A), we find that the statistics exhibit similar performance in small samples of 80 observations and they are not very powerful. A value of 0.05 is needed to give the MCCR statistic an edge so as to start outperforming. Also, the power increases in magnitude with correlation. As expected, for all statistics the power increases significantly with larger data sets. The MCCR statistic only slightly outperforms the MDM and the GW statistics when $b = 0.01$ but increasing the value of b results in a significant gain of power of the former relative to the latter two alternatives. For

[19] The power of a test statistic (i.e. 1 minus the probability of making a type II error) is by definition the probability of rejecting the null hypothesis when it is actually false.

[20] Higher values of b result in powers that are equal or very close to one.

Table 3 Rejection probabilities (%) of tests

Panel A: Empirical size of nominal 5% tests (DGP with normal errors)

		T = 80			T = 400			T = 1000			T = 1600		
		δ			δ			δ			δ		
Statistic	$\pi = P/R$	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	− 0.9
MDM	0.4	6.7	7.9	10.5	5.8	6.4	7.0	5.8	5.8	6.2	5.5	5.6	5.9
MCCR		5.9	6.7	8.2	5.4	5.7	5.6	5.7	5.5	5.4	5.6	5.5	5.5
GW		8.0	8.8	10.9	5.8	6.5	7.4	5.3	5.5	6.1	5.2	5.4	6.3
MDM	0.6	5.1	6.2	8.6	5.0	5.2	5.6	4.6	4.7	5.3	4.8	4.8	4.8
MCCR		4.7	5.5	6.7	4.6	4.9	5.0	4.6	4.7	4.9	4.8	4.8	4.7
GW		5.1	5.9	8.0	5.2	5.7	6.6	4.8	5.1	6.1	5.2	5.3	5.5
MDM	1.0	5.5	6.4	8.6	5.3	5.3	6.3	5.2	5.3	5.6	5.1	5.3	5.7
MCCR		5.2	5.6	7.3	4.9	4.9	5.4	5.1	4.8	5.1	5.0	5.1	5.3
GW		5.5	5.9	8.9	5.3	5.6	6.8	5.1	5.4	5.8	5.2	5.2	5.7
MDM	2.0	3.8	4.7	6.8	4.5	4.9	5.1	4.8	4.9	5.3	4.8	5.0	5.0
MCCR		4.1	4.9	6.5	4.7	5.0	5.0	4.9	4.9	5.2	4.8	5.1	5.0
GW		4.5	5.6	8.3	5.0	5.4	6.3	5.1	5.2	5.7	4.9	5.4	5.6

Panel B: Empirical size of nominal 5% tests (DGP with GARCH(1,1)-normal errors)

		T = 80			T = 400			T = 1000			T = 1600		
		δ			δ			δ			δ		
Statistic	$\pi = P/R$	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	− 0.9
MDM	0.4	6.7	7.5	10.6	6.0	6.0	7.0	5.6	5.8	6.1	5.4	5.7	5.8
MCCR		6.1	6.3	8.3	5.5	5.3	6.0	5.4	5.6	5.7	5.3	5.5	5.7
GW		7.9	8.6	11.1	6.1	6.1	7.1	5.2	5.4	6.4	5.1	5.5	6.2
MDM	0.6	5.4	6.3	8.4	5.0	5.4	6.2	4.6	4.9	5.3	4.7	4.8	4.9
MCCR		4.9	5.7	6.7	4.9	5.0	5.3	4.5	4.9	5.2	4.7	4.9	5.3
GW		5.6	6.1	8.0	5.1	5.9	6.7	4.8	5.1	5.8	4.7	5.1	5.7
MDM	1.0	5.2	6.4	8.8	5.2	5.8	6.6	5.3	5.5	5.8	5.2	5.5	5.3
MCCR		4.9	5.8	7.3	4.9	5.4	5.9	5.0	5.2	5.6	5.0	5.1	5.3
GW		5.3	6.4	8.9	5.2	5.9	6.7	5.1	5.5	5.7	5.2	5.2	5.5
MDM	2.0	3.9	5.1	7.0	4.6	4.8	5.3	4.7	5.1	5.1	5.0	4.7	5.2
MCCR		4.0	5.3	7.0	4.7	4.9	5.2	4.7	5.1	5.2	5.0	4.9	5.4
GW		4.2	5.7	8.6	5.1	5.3	6.2	5.1	5.2	5.6	5.1	5.1	5.5

Table 3 (continued)

Panel C: Empirical size of nominal 5% tests (DGP with GARCH(1,1)-*t*(5) errors)													
		T=80			T=400			T=1000			T=1600		
		δ			δ			δ			δ		
Statistic	$\pi = P/R$	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	− 0.9	0	− 0.5	− 0.9
MDM	0.4	7.3	8.9	10.7	6.1	6.6	7.5	5.6	5.5	6.1	5.4	5.5	6.1
MCCR		6.1	7.2	8.6	5.7	6.1	7.0	5.3	5.8	6.9	5.4	5.9	7.0
GW		7.6	8.8	11.2	5.7	6.5	7.6	5.1	5.3	6.2	5.0	5.4	6.0
MDM	0.6	5.6	7.2	9.2	5.3	5.6	6.1	4.8	5.1	5.4	4.6	5.0	5.1
MCCR		5.1	6.3	7.5	4.8	5.3	6.1	4.5	5.3	6.2	4.6	5.3	6.2
GW		5.5	6.8	8.3	5.2	5.9	6.8	4.8	5.6	6.2	4.8	5.3	5.7
MDM	1.0	5.7	7.0	9.2	5.5	5.8	6.4	5.2	5.4	6.0	5.4	5.2	5.8
MCCR		5.2	6.0	7.9	4.8	5.5	6.2	4.8	5.4	6.4	4.8	5.6	6.6
GW		5.4	6.8	8.8	5.3	5.8	6.8	5.2	5.4	6.2	5.0	5.0	5.6
MDM	2.0	4.4	5.6	7.1	4.8	5.4	5.7	4.7	4.9	5.6	5.0	5.1	4.9
MCCR		4.4	5.4	7.0	4.5	5.4	6.0	4.6	5.2	6.4	4.9	5.4	6.1
GW		4.8	6.3	8.6	4.7	5.5	6.5	4.8	5.3	5.8	5.1	5.2	5.3

Explanation: This table presents the empirical size of the graphical procedure of Goyal and Welch (2003) (GW) together with the empirical size of the more conventional tests of equal predictive accuracy, at the 5% nominal level. The number of replications is 25,000. MDM denotes the modified version of the Diebold and Mariano (1995) statistic suggested by Harvey et al. (1997) while MCCR denotes the test developed by McCracken (2007). We consider different proportions of out-of-sample (P) to in-sample (R) observations, denoted by π, as well as different levels of correlation (δ) between the errors of the returns and dividend-price ratio processes. Panel A reports results with respect to the returns data generating process (DGP) that is based on Gaussian errors while Panels B and C show the results with respect to the DGP that uses GARCH(1,1)-normal errors and GARCH(1,1)-*t* errors with five degrees of freedom, respectively (see Section 2.1 for a detailed description)

example, when $T = 1600$, $\delta = -0.9$, $\pi = 1.0$ and b is set from 0.01 to 0.05, the empirical rejection probability for MCCR increases from 15.6% to 97.3%. On the other hand, the MDM statistic increases its empirical rejection probability from 15.1% to 87.9% while the graphical procedure starts from 14.2% and yields 71.9%. Results are similar with respect to the GARCH(1,1)-normal case (Table 4, Panel B).

When we model the innovations of the returns series as a GARCH(1,1)-*t* process with five degrees of freedom, the MCCR statistic is still found to be the most powerful statistic (Table 4, Panel C). The only difference comes from the fact that all corresponding powers decrease in magnitude compared to the homoscedastic case.

To briefly summarize, we find that the GW graphical procedure exhibits a satisfactory size and power performance in finite samples and appears to be quite robust under most considered scenarios. Hence, in conjunction with its dynamic nature which may reveal useful information in an empirical study, our paper suggests that it can be an important complementary tool to more conventional methods seeking to evaluate out-of-sample predictive ability.

3.4 Using small-sample critical values

Finally, we obtained small-sample simulated critical values and repeated all size and power experiments for the graphical diagnostic of GW.[21] Consistent with the approach we followed to obtain large-sample critical values based on a sample of 6000 observations (explained in Sect. 2.4.3), we repeated the same steps for smaller samples of 80, 400, 1000 and 1600 observations. Based on 50,000 replications for each sample size, we calculated q_i^{jk}, as defined in Eq. (10), at each replication. For all cases, the 95th percentile of the empirical distribution is chosen as the critical value at the 5% level. The small-sample critical values can be found in Table 5 of the Appendix, whereas Table 6 and Table 7 present the size and power results, respectively, across all scenarios examined in our primary analysis.

As can be seen in Table 6, our results remain robust, and our main inferences are unaffected. In particular, we observe that the graphical diagnostic exhibits good size properties overall and correctly rejects the null under most scenarios. The rejection probabilities follow a similar pattern across the different data generation processes we consider. As expected, correlation occasionally has a greater impact on the smaller sample of 80 observations. Note that there are no tabulated results for $T = 80$ and $\pi = 0.4$, as the out-of-sample period in this case is too small to generate a critical value less than unity (see also Table 5).

Moving on to the power results which are tabulated in Table 7, we again draw similar conclusions. For the two extreme sample sizes of our study, we find that using the small-size simulated critical values leads to qualitatively similar results. In line with our previous findings, all powers increase as b increases from 0.01 to 0.05. As expected, the estimated powers are smaller for the sample size of 80 observations while there is a significant power increase with the larger sample of 1600 observations. Again, the power results reveal a similar pattern under different scenarios of data generating processes.

[21] We are grateful to an anonymous reviewer for this suggestion.

Table 4 Power results

Panel A: Empirical power of nominal 5% tests (DGP with normal errors)										
		$\delta = 0$			$\delta = -0.5$			$\delta = -0.9$		
			b			b			b	
Statistic	$\pi = P/R$	0.01	0.02	0.05	0.01	0.02	0.05	0.01	0.02	0.05
		T=80								
MDM	0.6	5.4	5.9	9.9	6.7	7.7	12.3	8.4	9.5	15.1
MCCR		5.3	6.1	12.4	6.3	7.7	14.9	7.6	9.0	18.3
GW		5.7	6.0	8.9	6.1	7.2	10.9	7.8	8.5	13.3
MDM	1.0	5.6	5.9	11.1	6.6	8.1	14.1	9.0	10.4	16.9
MCCR		5.3	6.0	12.9	6.3	7.9	16.2	8.3	10.0	19.5
GW		5.2	5.5	9.5	6.3	7.8	11.9	8.4	9.9	16.0
MDM	2.0	4.4	5.0	9.4	5.6	6.7	13.1	7.7	8.7	16.0
MCCR		4.6	5.6	11.6	5.9	7.3	15.4	7.5	9.4	18.6
GW		5.0	5.3	8.7	6.2	7.0	12.5	8.6	10.4	17.8
		T=1600								
MDM	0.6	10.2	25.1	69.4	11.3	26.2	71.7	12.1	26.4	73.5
MCCR		12.5	36.0	88.9	13.3	37.9	91.7	14.1	38.4	94.4
GW		9.0	19.6	45.4	11.0	22.6	51.3	13.0	26.2	58.9
MDM	1.0	12.7	31.6	83.4	13.7	33.1	85.3	15.1	34.5	87.9
MCCR		14.4	39.4	93.5	15.2	41.6	95.5	15.6	43.3	97.3
GW		10.8	24.0	59.7	12.7	27.3	65.7	14.2	31.5	71.9
MDM	2.0	13.4	37.0	92.2	14.3	38.1	93.8	14.9	39.2	95.8
MCCR		14.8	43.0	96.8	15.8	44.9	97.9	16.1	46.2	99.2
GW		11.7	28.7	75.0	13.4	32.1	79.6	14.7	35.1	84.3
Panel B: Empirical power of nominal 5% tests (DGP with GARCH(1,1)-normal errors)										
		$\delta = 0$			$\delta = -0.5$			$\delta = -0.9$		
			b			*b*			*b*	
Statistic	$\pi = P/R$	0.01	0.02	0.05	0.01	0.02	0.05	0.01	0.02	0.05
		T=80								
MDM	0.6	5.8	6.5	11.2	7.5	8.9	15.3	10.5	11.9	18.6
MCCR		5.6	6.3	14.2	6.9	8.5	18.4	9.0	11.0	22.6
GW		5.8	6.0	9.5	7.1	7.9	13.2	9.4	10.3	16.0
MDM	1.0	5.8	6.4	13.0	7.8	9.3	17.6	10.3	12.3	20.3
MCCR		5.5	6.5	15.4	7.0	8.9	20.0	9.2	12.0	23.5
GW		5.3	6.0	10.6	7.2	8.4	15.2	10.1	12.1	19.1
MDM	2.0	4.3	5.1	12.1	6.1	7.7	16.1	8.5	10.7	18.4
MCCR		4.6	5.6	14.5	6.3	8.2	18.9	8.7	11.3	21.9
GW		4.7	5.5	10.4	6.9	8.1	15.6	10.3	12.4	20.6
		T=1600								
MDM	0.6	12.0	27.4	73.9	12.8	28.7	75.6	13.4	29.2	76.7
MCCR		14.8	40.1	91.6	15.3	41.8	93.4	16.2	43.4	95.4
GW		10.1	21.4	47.5	12.4	25.2	55.2	14.3	29.1	62.9
MDM	1.0	14.3	35.3	87.0	15.0	36.3	88.3	16.0	36.9	89.5
MCCR		16.1	44.4	95.4	16.4	46.0	96.9	17.5	47.4	97.8
GW		11.8	26.5	63.3	13.8	30.5	69.6	15.3	33.4	75.0
MDM	2.0	14.7	41.1	94.7	15.6	42.5	95.8	16.7	42.4	96.3
MCCR		16.5	48.2	98.1	17.2	50.0	98.9	18.4	50.9	99.3
GW		12.4	32.2	78.8	14.1	35.3	82.9	15.8	38.1	86.1

Table 4 (continued)

Panel C: Empirical power of nominal 5% tests (DGP with GARCH(1,1)-*t*(5) errors)										
		$\delta = 0$			$\delta = -0.5$			$\delta = -0.9$		
			b			*b*			*b*	
Statistic	$\pi = P/R$	0.01	0.02	0.05	0.01	0.02	0.05	0.01	0.02	0.05
		T = 80								
MDM	0.6	5.8	6.0	8.5	7.9	8.5	12.2	9.5	11.1	15.7
MCCR		5.2	5.8	9.2	7.0	7.9	12.7	8.2	10.1	17.1
GW		5.8	5.9	7.4	7.4	7.9	10.6	8.8	9.8	13.8
MDM	1.0	6.0	6.1	8.8	8.0	8.4	13.0	10.3	11.6	17.1
MCCR		5.4	5.8	9.4	7.0	7.7	13.2	9.1	10.6	18.1
GW		5.6	6.1	7.7	7.4	7.8	12.0	10.2	11.3	16.1
MDM	2.0	4.4	4.7	7.7	6.5	7.5	11.8	8.4	9.5	15.0
MCCR		4.5	5.0	8.4	6.3	7.3	12.5	8.5	9.9	16.8
GW		4.9	5.0	7.3	7.3	7.8	11.9	9.9	11.1	16.9
		T = 1600								
MDM	0.6	6.8	12.8	40.7	8.3	15.2	44.2	9.2	18.0	48.5
MCCR		7.6	16.0	58.9	9.6	19.6	65.8	11.5	25.0	74.5
GW		6.6	11.4	29.9	8.5	14.7	37.2	10.3	18.2	44.8
MDM	1.0	7.8	15.6	51.7	9.5	18.1	54.8	54.8	20.8	61.0
MCCR		8.1	17.7	64.4	9.7	21.4	70.7	12.3	26.4	78.8
GW		7.2	12.9	38.9	8.9	16.2	45.9	10.6	19.9	53.5
MDM	2.0	8.0	16.6	60.6	9.7	19.9	64.1	10.4	22.5	70.0
MCCR		8.4	18.7	69.4	10.5	22.5	74.9	12.1	27.4	82.0
GW		7.4	14.5	48.1	8.9	17.8	54.2	10.8	21.1	62.6

Explanation: This table tabulates the empirical power of the graphical procedure of Goyal and Welch (2003) (GW) together with the empirical power of the more conventional tests of equal predictive accuracy, at the 5% nominal level. MDM denotes the modified version of the Diebold and Mariano (1995) statistic suggested by Harvey et al. (1997) while MCCR denotes the test developed by McCracken (2007). The empirical powers are in percentages and the number of simulations is 25,000. For brevity, power results are shown only for sample sizes of 80 and 1600 observations. We consider different proportions of out-of-sample (P) to in-sample (R) observations, denoted by π, as well as different levels of correlation (δ) between the errors of the returns and dividend-price ratio processes. Panel A reports results with respect to the returns data generating process (DGP) that is based on Gaussian errors while Panels B and C show the results with respect to the DGP that uses GARCH(1,1)-normal errors and GARCH(1,1)-*t* errors with five degrees of freedom, respectively (see Section 2.1 for a detailed description)

4 Conclusion

The present paper extends Goyal and Welch's (2003) (GW) work and uses Monte Carlo simulations to assess the robustness of the recursive residuals graphical approach which they propose as a powerful diagnostic for equity premium and stock return prediction. The GW graphical technique has received much attention in the literature and numerous studies employ it as a useful tool to assess the relative out-of-sample predictive performance of competing models in various contexts (for recent applications see, *inter alia*, Alexandridis et al. 2023; Deng et al. 2024; Liu et al. 2024; Yin and Yang 2024; Gao et al. 2025; Hong et al. 2025). We contribute to the extant financial econometrics and return predictability literature by simulating the graphical procedure and we offer new evidence by exploring its unknown finite-sample properties. This in an important issue which is of particular interest for both academics and investors who use only currently available data to construct forecasts and to

form investment strategies. We base our approach on a sign test which is derived from the algebraic representation of the diagnostic. Our study is motivated by the dynamic nature of the graphical method which allows us to identify the actual time periods where a predictive variable succeeds (or fails) in predicting stock returns. Hence, graphing recursive residuals may uncover valuable insights into stock return predictability, providing a strong rationale for further examination of this approach.

Turning to our methodology, we first obtain simulated recursive out-of-sample forecasts which are derived from two models: The unconditional historical moving average model (i.e. our benchmark model) and the conditional dividend-price ratio model. At each replication, the algebraic expression of the graphical diagnostic is computed. Since a positive value in the graph is an indication of out-of-sample predictive ability, we focus on the expression that represents the percentage of positive values. Our findings indicate that the presence of correlation affects the graph in small samples but the procedure is robust when sufficiently large data sets are employed and the impact of correlation diminishes.

Taking our research even further, we run experiments and examine the finite- sample size and power properties of the procedure. For that purpose, we calculate empirical critical values with respect to different proportions of out-of-sample to in-sample observations and we perform one-sided tests at the 5% significance level. As a result, our study offers a "formal" comparison between the graphical diagnostic and the more conventional tests of out-of-sample forecasting accuracy. In particular, we also consider the classic Diebold and Mariano (1995) (DM) statistic, its modified version suggested by Harvey et al. (1997) (MDM) as well as a subsequent and influential test developed by McCracken (2007) (MCCR) (for recent empirical applications of these tests, see Gupta et al. 2023; Lan et al. 2024; Yang et al. 2024; Labonne 2025).

With respect to size experiments, we find that in small samples all employed statistics are reasonably sized, given that the level of correlation is moderate. However, the tests become more oversized when a high correlation is considered. Under this scenario, the MCCR test appears to be relatively more reliable. As the sample size increases, the considered statistics are appropriately sized and exhibit a similar performance. Apart from the case of small data sets, the graphical diagnostic of GW is found to be well-sized and retains good size properties even under the assumption of heteroskedasticity in the errors which is a common feature of financial data.

Furthermore, power simulations reveal that all tests yield relatively low powers in small samples and they display a similar performance. On the other hand, when larger data sets are employed, the corresponding powers significantly increase in magnitude and all statistics achieve a solid performance. In line with previous work, the MCCR statistic is generally found to be the most powerful in a strictly statistical sense (see, McCracken 2007). The results follow a similar pattern for all cases of data generating processes under consideration (i.e. based on normal as well as heteroskedastic errors).

Overall, our simulations suggest that the graphical procedure of GW can be an important complement to more conventional methods seeking to assess out-of-sample predictive ability. Hence, given it can convey useful information and reveal hidden periods of out-of-sample predictive (in)ability, our paper further advocates its future use by academics and investors alike.

Appendix for "A graphical procedure for equity premium and stock return prediction: Monte Carlo evidence"

Table 5 Small-sample simulated empirical critical values

T = 80			
$\pi = \frac{P}{R}$	99th Percentile	95th Percentile	90th Percentile
0.1	1.000	1.000	1.000
0.2	1.000	1.000	0.923
0.4	1.000	1.000	0.913
0.6	1.000	0.967	0.833
0.8	1.000	0.944	0.806
1.0	1.000	0.925	0.775
1.2	1.000	0.886	0.727
1.4	1.000	0.891	0.717
1.6	1.000	0.857	0.673
1.8	1.000	0.843	0.647
2.0	1.000	0.811	0.623
3.0	0.967	0.717	0.517
T = 400			
$\pi = \frac{P}{R}$	99th Percentile	95th Percentile	90th Percentile
0.1	1.000	1.000	0.944
0.2	1.000	0.985	0.925
0.4	1.000	0.965	0.877
0.6	1.000	0.947	0.840
0.8	1.000	0.933	0.809
1.0	0.995	0.915	0.770
1.2	0.991	0.894	0.748
1.4	0.991	0.880	0.721
1.6	0.992	0.874	0.707
1.8	0.988	0.849	0.674
2.0	0.985	0.835	0.652
3.0	0.970	0.763	0.567
T = 1000			
$\pi = \frac{P}{R}$	99th Percentile	95th Percentile	90th Percentile
0.1	1.000	0.989	0.934
0.2	1.000	0.976	0.910
0.4	1.000	0.962	0.871
0.6	0.997	0.941	0.827
0.8	0.995	0.921	0.791
1.0	0.994	0.912	0.770
1.2	0.993	0.895	0.749
1.4	0.991	0.873	0.717
1.6	0.990	0.867	0.696
1.8	0.989	0.854	0.681
2.0	0.985	0.841	0.666
3.0	0.969	0.775	0.583

Table 5 (continued)

T = 1600			
$\pi = \frac{P}{R}$	99th Percentile	95th Percentile	90th Percentile
0.1	1.000	0.986	0.938
0.2	1.000	0.974	0.907
0.4	0.998	0.958	0.871
0.6	0.997	0.943	0.832
0.8	0.996	0.928	0.800
1.0	0.994	0.911	0.771
1.2	0.992	0.892	0.741
1.4	0.992	0.883	0.727
1.6	0.990	0.861	0.701
1.8	0.986	0.848	0.678
2.0	0.988	0.838	0.668
3.0	0.973	0.785	0.594

Explanation: This table tabulates the small-sample simulated critical values at the usual levels of significance, for the sign-based test inspired by the graphical procedure of Goyal and Welch (2003). The 99th, 95th and 90th percentiles are derived from experiments which use samples of 80, 400, 1000 and 1600 observations. The number of repetitions is 50,000. Section 2.4.2 describes the simulation approach. To obtain the small-sample critical values in this table, we follow the same steps as explained in Section 2.4.3 in relation to a large sample of 6,000 observations. The ratio of out-of-sample (P) to in-sample (R) observations is denoted by π. Table 5 shows the results with respect to various values of π

Table 6 Rejection probabilities (%) of tests

		T = 80			**T = 400**			**T = 1000**			**T = 1600**		
		δ			δ			δ			δ		
DGP	π = P/R	0	-0.5	-0.9	0	-0.5	-0.9	0	-0.5	-0.9	0	-0.5	-0.9
DGP1	0.4	-	-	-	4.4	5.0	5.7	4.6	5.1	5.4	5.2	5.2	5.7
DGP2		-	-	-	4.7	4.8	5.6	4.7	4.8	5.8	5.1	5.5	6.0
DGP3		-	-	-	4.8	5.8	6.1	4.9	5.8	6.2	5.2	6.0	6.5
DGP1	0.6	3.4	4.1	4.6	5.0	5.1	6.1	5.2	5.3	5.8	4.9	5.2	5.6
DGP2		3.5	3.9	4.5	4.8	4.9	6.2	5.0	5.7	6.3	4.9	5.2	5.4
DGP3		3.5	4.2	4.7	5.0	6.2	7.1	5.2	6.1	6.7	5.0	5.6	6.3
DGP1	1.0	4.2	5.3	7.1	4.9	5.0	6.6	4.6	5.0	5.5	5.0	5.0	5.4
DGP2		4.0	5.2	7.0	4.7	5.2	6.3	4.9	5.2	5.9	5.0	5.0	5.6
DGP3		4.3	5.9	7.5	4.9	6.2	6.9	5.3	5.7	6.5	5.0	5.7	6.0
DGP1	2.0	5.3	6.9	9.7	5.0	5.9	6.2	4.9	5.2	5.7	5.1	5.2	5.3
DGP2		5.2	6.7	9.6	5.0	5.4	6.3	5.0	5.2	5.3	5.1	4.8	5.6
DGP3		5.5	8.1	10.1	5.4	6.3	7.0	5.3	5.7	5.9	5.3	5.7	5.6

Explanation: This table presents the empirical size of the graphical procedure of Goyal and Welch (2003) (GW) at the 5% nominal level when using small-sample simulated critical values. The number of replications is 25,000. We consider different proportions of out-of-sample (P) to in-sample (R) observations, denoted by π, as well as different levels of correlation (δ) between the errors of the returns and dividend-price ratio processes. Results are reported with respect to different data generating processes (DGP) for stock returns: DGP1 is based on Gaussian errors; DGP2 employs GARCH(1,1)-normal errors; and DGP3 uses GARCH(1,1)-*t* errors with five degrees of freedom (see Sections 2.1 and 3.4 for a detailed description)

Table 7 Empirical power of nominal 5% tests

		δ = 0			δ = -0.5			δ = -0.9		
			b			b			b	
DGP	π = P/R	0.01	0.02	0.05	0.01	0.02	0.05	0.01	0.02	0.05
	T=80									
DGP1	0.6	3.6	3.8	5.8	4.2	4.8	7.2	5.0	5.5	8.4
DGP2		3.6	4.0	5.9	4.1	4.6	6.9	4.9	5.3	8.4
DGP3		3.8	4.9	10.3	4.7	5.6	9.6	5.1	6.0	8.9
DGP1	1.0	4.3	4.9	7.8	5.3	6.5	10.1	6.9	7.8	12.5
DGP2		4.4	4.9	7.8	5.3	6.2	10.2	6.7	7.8	12.3
DGP3		4.8	6.5	14.5	6.3	7.7	14.5	7.2	8.6	13.0
DGP1	2.0	5.8	6.2	10.1	7.5	8.3	14.5	10.3	12.0	19.7
DGP2		5.5	6.4	10.2	6.9	8.1	14.4	9.8	11.3	19.1
DGP3		4.9	5.0	7.3	7.3	7.8	11.9	9.9	11.1	16.9
	T=1600									
DGP1	0.6	9.3	19.5	45.2	11.3	23.2	51.6	13.1	26.4	59.9
DGP2		9.5	19.4	45.3	11.1	22.5	51.4	12.9	26.2	58.6
DGP3		18.1	35.1	63.5	16.3	32.1	67.6	12.9	25.0	59.3
DGP1	1.0	10.7	24.0	58.0	12.1	27.0	64.4	13.8	30.7	72.0
DGP2		10.4	23.7	58.1	12.2	27.2	64.7	13.7	30.0	70.5
DGP3		22.2	45.4	78.3	17.3	39.0	80.7	13.5	28.3	71.0
DGP1	2.0	11.6	29.3	75.7	12.8	32.2	80.2	14.5	34.7	84.8
DGP2		12.1	28.8	75.0	12.8	31.9	79.7	14.1	34.1	83.6
DGP3		26.1	58.9	92.8	19.3	47.4	92.8	13.7	31.4	81.6

Explanation: This table tabulates the empirical power of the graphical procedure of Goyal and Welch (2003) (GW) at the 5% nominal level when using small-sample simulated critical values. The empirical powers are in percentages and the number of simulations is 25,000. In line with our primary analysis, power results are shown only for sample sizes of 80 and 1600 observations. We consider different proportions of out-of-sample (P) to in-sample (R) observations, denoted by π, as well as different levels of correlation (δ) between the errors of the returns and dividend-price ratio processes. Results are reported with respect to different data generating processes (DGP) for stock returns: DGP1 is based on Gaussian errors; DGP2 employs GARCH(1,1)-normal errors; and DGP3 uses GARCH(1,1)-*t* errors with five degrees of freedom (see Sections 2.1 and 3.4 for a detailed description)

Acknowledgments We sincerely thank the editor, Cheng-Few Lee, and the two anonymous reviewers for their insightful comments and constructive suggestions, which enabled us to improve our work. This paper is dedicated to the memory of John C. Nankervis, an exceptional mentor and distinguished academic.

Declarations

Conflict of interest The authors declare that there are no relevant financial or non-financial interests to disclose.

References

Adediran IA, Swaray R (2023) Carbon trading amidst global uncertainty: the role of policy and geopolitical uncertainty. Econ Model 123:106279

Alexandridis AK, Apergis I, Panopoulou E, Voukelatos N (2023) Equity premium prediction: the role of information from the options market. J Financ Mark 64:100801

Algaba A, Boudt K (2017) Generalized financial ratios to predict the equity premium. Econ Model 66:244–257

Amihud Y, Hurvich CM (2004) Predictive regressions: a reduced-bias estimation method. J Financ Quant Anal 39(4):813–841

Amihud Y, Hurvich C, Wang Y (2004) Hypothesis testing in predictive regressions. Working paper, NYU Finance Dept

Amini S, Elmore R, Öztekin Ö, Strauss J (2021) Can machines learn capital structure dynamics? J Corp Financ 70:102073

Andriosopoulos D, Chronopoulos DK, Papadimitriou FI (2014) Can the information content of share repurchases improve the accuracy of equity premium predictions? J Empir Finance 26:96–111

Ang A, Bekaert G (2007) Stock return predictability: is it there? Rev Financ Stud 20(3):651–707

Angelidis T, Sakkas A, Tessaromatis N (2025) Predicting commodity returns: time series vs. cross sectional prediction models. J Commodity Mark 38:100475

Asgharian H, Christiansen C, Hou AJ (2023) The effect of uncertainty on stock market volatility and correlation. J Bank Finance 154:106929

Basistha A (2023) Estimation of short-run predictive factor for US growth using state employment data. J Forecast 42(1):34–50

Bennett D, Mekelburg E, Strauss J, Williams TH (2024) Unlocking the black box of sentiment and cryptocurrency: what, which, why, when and how? Glob Financ J 60:100945

Berisha E, Gabauer D, Gupta R, Lau CKM (2021) Time-varying influence of household debt on inequality in United Kingdom. Empir Econ 61:1917–1933

Bjørnland HC, Ravazzolo F, Thorsrud LA (2017) Forecasting GDP with global components: this time is different. Int J Forecast 33(1):153–173

Bouri E, Gkillas K, Gupta R, Pierdzioch C (2021) Forecasting power of infectious diseases-related uncertainty for gold realized variance. Financ Res Lett 42:101936

Caldeira JF, Torrent H (2017) Forecasting the US term structure of interest rates using nonparametric functional data analysis. J Forecast 36(1):56–73

Caldeira JF, Moura GV, Santos AA (2016) Predicting the yield curve using forecast combinations. Comput Stat Data Anal 100:79–98

Campbell JY, Shiller RJ (1988a) Stock prices, earnings, and expected dividends. J Finance 43(3):661–676

Campbell JY, Shiller RJ (1988b) The dividend-price ratio and expectations of future dividends and discount factors. Rev Financ Stud 1(3):195–228

Campbell JY, Thompson SB (2008) Predicting the equity premium out-of-sample: can anything beat the historical average? Rev Financ Stud 21(4):1509–1531

Campbell JY, Yogo M (2006) Efficient tests of stock return predictability. J Financ Econ 81(1):27–60

Campisi G, Muzzioli S, De Baets B (2024) A comparison of machine learning methods for predicting the direction of the US stock market on the basis of volatility indices. Int J Forecast 40(3):869–880

Charles A, Darné O, Kim JH (2017) International stock return predictability: evidence from new statistical tests. Int Rev Financ Anal 54:97–113

Chen J, Jiang F, Li H, Xu W (2016) Chinese stock market volatility and the role of US economic variables. Pac Basin Finance J 39:70–83

Chen Q, Han Y, Huang Y, Jiang GJ (2025) Jump risk implicit in options market. J Financ Econom 23(2):nbaf002

Choi Y, Jacewitz S, Park JY (2016) A reexamination of stock return predictability. J Econom 192(1):168–189

Christoffersen PF, Diebold FX (2006) Financial asset returns, direction-of-change forecasting, and volatility dynamics. Manag Sci 52(8):1273–1287

Ciner C (2025) Forecasting the aggregate market volatility by boosted neural networks. Financ Res Lett 72:106505

Clark TE, West KD (2006) Using out-of-sample mean squared prediction errors to test the Martingale difference hypothesis. J Econom 135(1–2):155–186

Costantini M, Kunst RM (2021) On using predictive-ability tests in the selection of time-series prediction models: a Monte Carlo evaluation. Int J Forecast 37(2):445–460

Deng Y, Wang Y, Zhou T (2024) Macroeconomic expectations and expected returns. Forthcoming, J Financ Quant Anal

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263

Ellwanger R, Snudden S (2023) Forecasts of the real price of oil revisited: do they beat the random walk? J Bank Finance 154:106962
Engstrom E (2003) The conditional relationship between the equity risk premium and the dividend price ratio, Working Paper, Columbia Business School
Fama EF, French KR (1988) Dividend yields and expected stock returns. J Financ Econ 22(1):3–25
Félix L, Kräussl R, Stork P (2020) Implied volatility sentiment: a tale of two tails. Quant Finance 20(5):823–849
Ferson WE, Sarkissian S, Simin TT (2003) Spurious regressions in financial economics? J Finance 58(4):1393–1413
Foroni C, Ravazzolo F, Sadaba B (2018) Assessing the predictive ability of sovereign default risk on exchange rate returns. J Int Money Finance 81:242–264
Gao S, Zhang Z, Wang Y, Zhang Y (2023) Forecasting stock market volatility: the sum of the parts is more than the whole. Finance Res Lett 55(A):103849
Gao S, Wang S, Wang Y, Zhang Q (2025) ChatGPT and commodity return. J Futures Mark 45(3):161–175
Gkillas K, Gupta R, Pierdzioch C (2021) Forecasting realized volatility of bitcoin returns: tail events and asymmetric loss. Eur J Finance 27(16):1626–1644
Goetzmann WN, Jorion P (1993) Testing the predictive power of dividend yields. J Finance 48(2):663–679
Goyal A, Welch I (2003) Predicting the equity premium with dividend ratios. Manag Sci 49(5):639–654
Goyal A, Welch I, Zafirov A (2024) A comprehensive 2022 look at the empirical performance of equity premium prediction. Rev Financ Stud 37(11):3490–3557
Gupta R, Ji Q, Pierdzioch C, Plakandaras V (2023) Forecasting the conditional distribution of realized volatility of oil price returns: the role of skewness over 1859 to 2023. Finance Res Lett 58(C):104501
Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared errors. Int J Forecast 13(2):281–291
Harvey DI, Leybourne SJ, Taylor AR (2023) Improved tests for stock return predictability. Econ Rev 42(9–10):834–861
He C, Teräsvirta T (1999) Properties of moments of a family of GARCH processes. J Econ 92(1):173–192
Hjalmarsson E (2011) New methods for inference in long-horizon regressions. J Financ Quant Anal 46(3):815–839
Hong Y, Jiang F, Meng L, Xue B (2025) Forecasting inflation using economic narratives. J Bus Econ Stat 43(1):216–231
Inoue A, Killian L (2004) In-sample or out-of-sample tests of predictability: which one should we use? Econ Rev 23(4):371–402
Jordan SJ, Vivian AJ, Wohar ME (2014) Forecasting returns: new European evidence. J Empir Finance 26:76–95
Kellard NM, Nankervis JC, Papadimitriou FI (2010) Predicting the equity premium with dividend ratios: reconciling the evidence. J Empir Finance 17(4):539–551
Kostakis A, Magdalinos T, Stamatogiannis MP (2015) Robust econometric inference for stock return predictability. Rev Financ Stud 28(5):1506–1553
Kuntz LC (2020) Beta dispersion and market timing. J Empir Finance 59:235–256
Labonne P (2025) Asymmetric uncertainty: nowcasting using skewness in real-time data. Int J Forecast 41(1):229–250
Lamont O (1998) Earnings and expected returns. J Finance 53(5):1563–1587
Lan W, Lei B, Feng L, Tsai CL (2024) Maximum-subsampling test of equal predictive ability. J Bus Econ Stat 42(4):1344–1355
Lawrenz J, Zorn J (2017) Predicting international stock returns with conditional price-to-fundamental ratios. J Empir Finance 43:159–184
Lettau M, Ludvigson S (2001) Consumption, aggregate wealth and expected stock returns. J Finance 56(3):815–849
Lewellen J (2004) Predicting returns with financial ratios. J Financ Econ 74(2):209–235
Li D, Zhang F, Li X (2022) Can US trade policy uncertainty help in predicting stock market excess return? Finance Res Lett 49:103136
Lima LR, Meng F (2017) Out-of-sample return predictability: a quantile combination approach. J Appl Econom 32(4):877–895
Liu L, Hao X, Wang Y (2024) Solving the forecast combination puzzle using double shrinkages. Oxf Bull Econ Stat 86(3):714–741
Luo Q, Ma F, Wang J, Wu Y (2024) Changing determinant driver and oil volatility forecasting: a comprehensive analysis. Energy Econ 129:107187
McCracken MW (2007) Asymptotics for out-of-sample tests of Granger causality. J Econom 140(2):719–752
McMillan DG (2003) Non-linear predictability of UK stock market returns. Oxf Bull Econ Stat 65(5):557–573

Monticini A, Ravazzolo F (2014) Forecasting the intraday market price of money. J Empir Finance 29:304–315
Neely CJ, Rapach DE, Tu J, Zhou G (2014) Forecasting the equity risk premium: the role of technical indicators. Manag Sci 60(7):1772–1791
Nelson CR, Kim MJ (1993) Predictable stock returns: the role of small sample bias. J Finance 48(2):641–661
Nygaard K, Sørensen LQ (2024) Betting on war? Oil prices, stock returns and extreme geopolitical events. Energy Econ 136:107659
Park D, Hahn J, Eom YH (2024) Predicting the equity premium with financial ratios: a comprehensive look over a long period in Korea. Pac Basin Finance J 84:102320
Pesaran MH, Timmermann A (2000) A recursive modelling approach to predicting UK stock returns. Econ J 110(460):159–191
Pettenuzzo D, Ravazzolo F (2016) Optimal portfolio choice under decision-based model combinations. J Appl Econom 31(7):1312–1332
Pontiff J, Schall LD (1998) Book-to-market ratios as predictors of market returns. J Financ Econ 49(2):141–160
Procasky WJ, Yin A (2022) Forecasting high-yield equity and CDS index returns: does observed cross-market informational flow have predictive power? J Futures Mark 42(8):1466–1490
Procasky WJ, Yin A (2023) The impact of COVID-19 on the relative market efficiency and forecasting ability of credit derivative and equity markets. Int Rev Financ Anal 90:102926
Rapach DE, Wohar ME (2006) In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. J Empir Finance 13(2):231–247
Rapach DE, Zhou G (2022) Asset pricing: Time-series predictability. Oxford Research Encyclopedia of Economics and Finance: 1 – 34
Rapach DE, Strauss JK, Zhou G (2010) Out-of-sample equity premium prediction: combination forecasts and links to the real economy. Rev Financ Stud 23(2):821–862
Rapach DE, Strauss JK, Zhou G (2013) International stock return predictability: what is the role of the United States? J Finance 68(4):1633–1662
Robertson D, Wright S (2006) Dividends, total cash flow to shareholders, and predictive return regressions. Rev Econ Stat 88(1):91–99
Rozeff M (1984) Dividend yields are equity risk premiums. J Portf Manag 11(1):68–75
Sakkas A, Tessaromatis N (2022) Forecasting the long-term equity premium for asset allocation. Financ Anal J 78(3):9–29
Salisu AA, Gupta R, Cepni O, Caraiani P (2024) Oil shocks and state-level stock market volatility of the United States: a GARCH-MIDAS approach. Rev Quant Fin Acc 63(4):1473–1510
Schrimpf A (2010) International stock return predictability under model uncertainty. J Int Money Finance 29(7):1256–1282
Stauskas O, Westerlund J (2022) Tests of equal forecasting accuracy for nested models with estimated CCE factors. J Bus Econ Stat 40(4):1745–1758
Stöckl S, Kaiser L (2021) Higher moments matter! Cross-sectional (higher) moments and the predictability of stock returns. Rev Financ Econ 39(4):455–481
Su H, Ying C, Zhu X (2022) Disaster risk matters in the bond market. Finance Res Lett 47(A):102764
Tsiakas I, Li J, Zhang H (2020) Equity premium prediction and the state of the economy. J Empir Finance 58:75–95
Wang Y, Huang X, Huang Z (2024) Energy-related uncertainty and Chinese stock market returns. Finance Res Lett 62(B):105215
Welch I, Goyal A (2008) A comprehensive look at the empirical performance of equity premium prediction. Rev Financ Stud 21(4):1455–1508
West KD (1996) Asymptotic inference about predictive ability. Econometrica 64(5):1067–1084
Xie H (2019) Financial volatility modeling: the feedback asymmetric conditional autoregressive range model. J Forecast 38(1):11–28
Xu Y, Lien D (2022) Forecasting volatilities of oil and gas assets: a comparison of GAS, GARCH, and EGARCH models. J Forecast 41(2):259–278
Yang Y, Guo JE, Li Y, Zhou J (2024) Forecasting day-ahead electricity prices with Spatial dependence. Int J Forecast 40(3):1255–1270
Yin A (2019) Out-of-sample equity premium prediction in the presence of structural breaks. Int Rev Financ Anal 65:101385
Yin A (2021) Forecasting the market equity premium: does nonlinearity matter? Int J Econ Financ 13(5):1–9
Yin A (2022) Does the kitchen-sink model work forecasting the equity premium? Int Rev Financ 22(1):223–247
Yin X, Yang G (2024) Instantaneous volatility of the yield curve, variance risk premium and bond return predictability. J Empir Finance 77:101490

Yin L, Feng J, Liu L, Wang Y (2019) It's not that important: the negligible effect of oil market uncertainty. Int Rev Econ Finance 60:62–84

Zakamulin V (2017) Secular mean reversion and long-run predictability of the stock market. Bull Econ Res 69(4):E66–E93

Zhang Y, Ma F, Wang Y (2019) Forecasting crude oil prices with a large set of predictors: can LASSO select powerful predictors? J Empir Finance 54:97–117